Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

External validation of the ARISCAT, Gupta Respiratory Failure, and Gupta Pneumonia risk scores in patients undergoing open abdominal aortic surgery: a single-center retrospective cohort study

Katarzyna W. Jasińska
Jagiellonian University
Kamil Polok
Jagiellonian University
Maciej Chwała
St. John Grande Hospital
Krzysztof Sładek
Jagiellonian University
Wojciech Szczeklik
Jagiellonian University
Dorota Studzińska
dstudzinska@gmail.com

Jagiellonian University

Research Article

Keywords: ARISCAT, external validation, Gupta score, postoperative pulmonary complications, risk prediction

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11365654/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

The ARISCAT, Gupta Respiratory Failure, and Gupta Pneumonia scores predict postoperative pulmonary complications, but their validity in open abdominal aortic surgery — a high-risk population — is unestablished. We aimed to externally validate all three scores in this population.

Methods.

Single-center retrospective cohort study of 693 adults undergoing open abdominal aortic surgery for infrarenal aortic aneurysm and/or peripheral arterial and aortic disease, at a tertiary center in Krakow, Poland, 2015–2023. Discrimination (area under the receiver operating characteristic curve) and calibration (calibration-in-the-large, slope, observed-to-expected ratio, calibration belt) of the three scores were assessed for their respective endpoints — composite pulmonary complication, respiratory failure, and pneumonia. Net clinical benefit was assessed by decision curve analysis, per the TRIPOD + AI 2024 statement.

Results.

The cohort was predominantly male (85.1%), median age 68 years (interquartile range, 64–73), and high surgical risk (73.0% American Society of Anesthesiologists physical status III). Postoperative pulmonary complication incidence was 19.5% (n = 135), respiratory failure 8.7% (n = 60), and pneumonia 9.1% (n = 63). All three scores showed poor discrimination: ARISCAT area under the receiver operating characteristic curve 0.563 (95% confidence interval, 0.512–0.615), Gupta Respiratory Failure 0.573 (0.516–0.631), Gupta Pneumonia 0.604 (0.536–0.672). All scores systematically underestimated risk (calibration-in-the-large, + 0.426, + 0.369, and + 0.837; observed-to-expected ratios, 1.38, 1.40, and 2.17; calibration belt $P < 0.001$, $P = 0.010$, and $P < 0.001$). Originally proposed cut-offs classified 75%–83% of patients as high-risk. Decision curve analysis showed net benefit over a treat-all strategy only for ARISCAT, within a narrow band near a 20% threshold; the Gupta scores showed no net benefit at any threshold.

Conclusions.

The ARISCAT, Gupta Respiratory Failure, and Gupta Pneumonia scores provide insufficient pulmonary risk stratification for patients undergoing open abdominal aortic surgery. Systematic underestimation of risk, combined with limited net clinical benefit at their original cut-offs, underscores the need for an aortic-specific pulmonary risk prediction model.

Trial registration.

Open Science Framework (OSF), https://doi.org/10.17605/OSF.IO/349UR. Registered 9 June 2026.

Background

Postoperative pulmonary complications (PPCs) are among the most common and clinically meaningful adverse events after non-cardiac surgery, with reported incidences ranging from approximately 5%-40% depending on population and definition [1, 2]. They prolong hospital stay, increase intensive care unit admission rate, and independently raise short-term mortality [1, 2]. In vascular and especially open abdominal aortic surgery (OAAS), where patients typically carry a substantial burden of comorbidities, perioperative physiological stress makes the lungs a particularly vulnerable organ in the perioperative period.

Several preoperative scores have been developed to quantify PPC risk. The ARISCAT score, derived by Canet and colleagues in 2010, uses seven routinely available preoperative variables to stratify patients into low, intermediate, and high-risk categories for a composite PPC endpoint [3]. The Gupta Postoperative Respiratory Failure (Gupta RF) score and the Gupta Postoperative Pneumonia score were both derived from a large national surgical quality improvement database, predicting their respective endpoints using preoperative variables [4, 5].

All three scores were derived from heterogeneous surgical populations in which open abdominal aortic procedures formed only a small subgroup [3–5]. External validations of ARISCAT score have been published across surgical specialties, with mixed results [6, 7]. To our knowledge, however, no study has specifically evaluated these three instruments in a large cohort of patients undergoing OAAS — a population with uniformly high baseline pulmonary risk, in which a well-calibrated preoperative tool would have direct clinical value. This is particularly relevant for internists performing preoperative risk stratification before referral for major vascular surgery. Current society guidelines for abdominal aortic aneurysm (AAA) repair likewise base pulmonary risk assessment on simple clinical criteria — symptomatic chronic obstructive pulmonary disease (COPD), long-standing tobacco use, or reduced exercise capacity — as an indication for pulmonary function testing (Level of recommendation 2/Weak; Quality of evidence C/Low), rather than on a validated multivariable prediction model [8].

We therefore externally validated the ARISCAT, Gupta RF, and Gupta Pneumonia scores in a retrospective cohort of patients undergoing OAAS due to infrarenal AAA and/or peripheral arterial and aortic diseases (PAAD).

Methods

Study design and population

We conducted a single-center retrospective cohort study at the Department of Vascular Surgery of St. John's Grande Hospital in Krakow, Poland — a tertiary vascular referral unit. A total of 923 patients aged ≥ 18 years undergoing major OAAS between January 2011 and December 2023 were screened for inclusion. 230 were excluded due to incomplete data precluding calculation of at least one risk score (all operated before 2015). The resulting complete-case cohort of 693 patients comprised exclusively those who had undergone OAAS due to infrarenal AAA and/or PAAD. Ethical approval for this study (No. OIL/KBL/13/2016) was provided by the Bioethics Committee of the District Medical Chamber in Krakow, Poland; the Committee waived the requirement for informed consent because of the retrospective design. The study was conducted in accordance with the latest version of the Declaration of Helsinki. The study protocol was registered on OSF (https://doi.org/10.17605/OSF.IO/349UR). This manuscript adheres to the applicable TRIPOD + AI 2024 guidelines [9].

Data sources and predictors

Preoperative, intraoperative, and postoperative data were extracted from the hospital electronic and paper medical record by trained investigators using a standardized case-report form; ambiguous outcomes were adjudicated by DS. Details are provided in Table 1 and Additional file 1: Table S1. All variables required to compute ARISCAT, Gupta RF, and Gupta Pneumonia were available for the analytic cohort with no missing data within the score components themselves. Sex was determined as recorded in the hospital medical record at the time of clinical care. To ensure participant privacy, data were pseudonymized upon completion of data collection and prior to statistical analysis by assigning IDs to each record. As this was a retrospective analysis, formal blinding of investigators to patient outcomes was not applicable. However, the definitions of predictor variables (ARISCAT, Gupta RF, and Gupta Pneumonia score components) and outcome events were fixed according to the original, previously published score definitions and standard clinical criteria, and were not modified based on the study findings.

Table 1
Baseline characteristics of the study population by postoperative pulmonary complication status (n = 693).

Variable	Overall (n = 693)	No PPC (n = 558)	PPC (n = 135)	P	Missing, n (%)
Demographics					
Age, years, median [IQR]	68.0 [64.0, 73.0]	68.0 [63.0, 73.0]	71.0 [66.0, 75.0]	**< 0.001**	0
Male sex, n (%)	590 (85.1)	480 (86.0)	110 (81.5)	0.232	0
BMI, kg/m², median [IQR]	26.7 [23.9, 29.4]	26.8 [23.8, 29.3]	26.4 [24.3, 30.1]	0.541	85 (12.3)
Comorbidities					
Hypertension, n (%)	559 (80.7)	448 (80.3)	111 (82.2)	0.697	0
Coronary artery disease, n (%)	184 (26.6)	145 (26.0)	39 (28.9)	0.564	0
COPD, n (%)	90 (13.0)	55 (9.9)	35 (25.9)	**< 0.001**	0
Congestive heart failure, n (%)	49 (7.1)	35 (6.3)	14 (10.4)	0.141	0
Cerebrovascular event (prior), n (%)	44 (6.3)	36 (6.5)	8 (5.9)	0.978	0
Smoking status, n (%)				0.882	0
Never	100 (14.5)	79 (14.2)	21 (15.7)		
Current	352 (50.9)	283 (50.8)	69 (51.5)		
Former	241 (34.8)	196 (35.1)	45 (33.3)		
Preoperative status					
Functional independence (complete), n (%)	687 (99.1)	554 (99.3)	133 (98.5)	0.332	0
Preoperative SpO_2, %, median [IQR]	97.0 [96.0, 98.0]	97.0 [96.0, 98.0]	97.0 [95.0, 98.0]	0.103	0
ASA physical status, n (%)				**0.031**	0
II [a]	144 (20.8)	127 (22.8)	17 (12.6)		
III [a]	506 (73.0)	398 (71.3)	108 (80.0)		
IV	43 (6.2)	33 (5.9)	10 (7.4)		
Procedure characteristics					

Variable	Overall (n = 693)	No PPC (n = 558)	PPC (n = 135)	P	Missing, n (%)
Demographics					
Procedure type, n (%)				0.265	0
OAR with tube prosthetic graft	371 (53.5)	298 (53.4)	73 (54.1)		
OAR with bifurcated prosthetic graft	194 (28.0)	151 (27.1)	43 (31.9)		
Isolated aortic bypass (due to aortoiliac occlusive disease)	128 (18.5)	109 (19.5)	19 (14.1)		
Anesthesia type, n (%)				**0.001**	0
General anesthesia alone	432 (62.3)	330 (59.1)	102 (75.6)		
General + epidural	261 (37.7)	228 (40.9)	33 (24.4)		
Procedure urgency, n (%)				0.163	0
Elective	643 (92.8)	522 (93.5)	121 (89.6)		
Expedited	50 (7.2)	36 (6.5)	14 (10.4)		
Operative duration, min, median [IQR]	150.0 [120.0, 180.0]	145.0 [116.2, 180.0]	155.0 [125.0, 190.0]	0.074	0
Aortic cross-clamp time, min, median [IQR]	55.0 [40.0, 70.0]	55.0 [40.0, 70.0]	55.0 [40.0, 70.0]	0.454	98 (14.1)
Preoperative risk scores					
ARISCAT score, median [IQR]	34.0 [34.0, 41.0]	34.0 [31.0, 41.0]	34.0 [34.0, 41.5]	**0.018**	0
ARISCAT risk category, n (%)				0.139	0
Low (≤ 25)	121 (17.5)	105 (18.8)	16 (11.9)		
Intermediate (26–44)	503 (72.6)	400 (71.7)	103 (76.3)		
High (≥ 45)	69 (10.0)	53 (9.5)	16 (11.9)		
Gupta RF predicted risk, %, median [IQR]	6.7 [6.7, 6.7]	6.7 [6.7, 6.7]	6.7 [6.7, 6.7]	**0.010**	0
Gupta Pneumonia predicted risk, %, median [IQR]	4.0 [3.2, 5.2]	3.9 [3.0, 5.2]	4.9 [3.6, 5.7]	**< 0.001**	0
RCRI score, median [IQR]	1.0 [1.0, 2.0]	1.0 [1.0, 2.0]	1.0 [1.0, 2.0]	0.321	0

Variable	Overall (n = 693)	No PPC (n = 558)	PPC (n = 135)	P	Missing, n (%)
Demographics					
P-POSSUM score, median [IQR]	4.4 [3.1, 6.0]	4.0 [3.1, 6.0]	5.0 [3.7, 7.0]	**< 0.001**	21 (3.0)

Table 3
Discrimination (AUC, 95% CI) of the scores for the three primary endpoints (n = 693).

Endpoint (events)	ARISCAT	Gupta RF	Gupta Pneumonia
PPC composite (n = 135)	**0.563 (0.512–0.615)**	0.556 (0.514–0.594)	0.628 (0.574–0.678)
Respiratory failure (n = 60)	0.604 (0.534–0.674)	**0.573 (0.516–0.631**)	0.632 (0.566–0.698)
Pneumonia (n = 63)	0.568 (0.501–0.635)	0.561 (0.500–0.622)	**0.604 (0.536–0.672)**

For each primary endpoint, the score whose original derivation outcome is bolded; 95% CI calculated by the DeLong method. The DeLong test identified a significant difference only between Gupta Pneumonia and Gupta RF for the respiratory failure endpoint (P = 0.020), whereas no other pairs of scores showed significant differences across any primary endpoints (P > 0.05). Abbreviations: AUC, area under the receiver operating characteristic curve; ARISCAT, Assess Respiratory Risk in Surgical Patients in Catalonia score; Gupta RF, Gupta Postoperative Respiratory Failure score.

Patients are stratified by occurrence of the postoperative pulmonary complication (PPC) composite endpoint. P-values from Mann–Whitney U test (continuous variables), Pearson χ^2 test (categorical variables), or Fisher's exact test where expected cell counts were below 5 (functional status). For variables with more than two categories reaching p < 0.05 overall, pairwise post-hoc comparisons were performed with Bonferroni correction. [a] ASA II vs III: p = 0.035 (Bonferroni-corrected); ASA II vs IV: p = 0.245; ASA III vs IV: p = 1.000. Bold values indicate p < 0.05. Abbreviations: ARISCAT, Assess Respiratory Risk in Surgical Patients in Catalonia; ASA, American Society of Anesthesiologists; BMI, body mass index; COPD, chronic obstructive pulmonary disease; IQR, interquartile range; OAR, open aneurysm repair; P-POSSUM, Portsmouth Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity; PPC, postoperative pulmonary complication; RCRI, Revised Cardiac Risk Index.

ARISCAT, Gupta RF, and Gupta Pneumonia scores were calculated according to the original published definitions of Canet and colleagues [3], Gupta and colleagues (respiratory failure) [4], and Gupta and colleagues (pneumonia) [5], respectively, and constituted the primary scores under validation. Procedure urgency was coded into three categories — elective, expedited (time-sensitive), and emergency.

Outcomes

The primary outcome for the ARISCAT validation was the composite PPC endpoint defined in the original derivation study — at least one of: respiratory failure, respiratory infection (pneumonia), pleural effusion, atelectasis, pneumothorax, bronchospasm, or aspiration pneumonitis, occurring during the index hospital admission [3]. The primary outcome for the Gupta RF validation was the composite of unplanned reintubation within 30 days or mechanical ventilation longer than 48 h [4]. The primary outcome for the Gupta Pneumonia validation was postoperative pneumonia, defined as a clinical diagnosis documented in the medical record.

Secondary outcomes were the individual PPC components, prolonged mechanical ventilation longer than 8 h, postoperative myocardial infarction, new-onset atrial fibrillation, pulmonary edema, acute kidney injury (Kidney Disease: Improving Global Outcomes [KDIGO] criteria), renal replacement therapy (RRT), sepsis, in-hospital mortality, and 30-day mortality.

Statistical analysis

Continuous variables were summarized as median [interquartile range, IQR] and compared with the Mann–Whitney U test, as normality could not be assumed (Shapiro–Wilk test, $P < 0.05$ for the majority of variables). Categorical variables were summarized as n (%) and compared with the chi-squared test, or Fisher's exact test when expected cell counts were below 5. Post-hoc pairwise comparisons used Bonferroni correction where applicable.

Discrimination was quantified by the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals calculated by the DeLong method [10]. Paired AUC comparisons between scores on the same outcome were performed with the DeLong test [10].

Calibration was assessed using four complementary approaches: (i) calibration-in-the-large (CITL); (ii) calibration slope; (iii) calibration belt (givitiR package) [11]; and (iv) the Hosmer–Lemeshow test, reported for completeness [12, 13]. All calibration confidence intervals used 2000 bootstrap resamples.

Original published predicted probabilities were used without refitting. Because only three (ARISCAT) and four (Gupta RF) unique predicted probabilities exist in this cohort, CITL, observed-to-expected (O/E) ratio, and calibration belt were the primary calibration metrics; slope and Hosmer–Lemeshow test are reported for completeness. Decision curve analysis (DCA) was performed at thresholds of 5–50% (PPC endpoint) and 2–30% (respiratory failure and pneumonia). Youden-optimized and developer-proposed cut-offs were evaluated with Wilson 95% confidence intervals (CIs). All tests were two-sided ($P < 0.05$). Analyses used R version 4.5.3 (R Foundation for Statistical Computing, Vienna, Austria) (pROC, givitiR, dcurves, rms).

Primary outcomes were the original derivation endpoints of each score and were fixed before analysis; the study protocol was registered retrospectively on OSF. Cut-off analyses and secondary outcomes were exploratory.

Given the retrospective design, no formal a priori sample size calculation was performed; the cohort comprised all eligible patients treated during the study period (a consecutive, all-comers sample).

Results

Study population

Between January 2015 and December 2023, 693 patients undergoing OAAS met the inclusion criteria and were enrolled in the analytic cohort. Baseline characteristics, stratified by the occurrence of the composite PPC endpoint, are presented in Table 1 and Additional file 1: Table S1. The cohort was predominantly male (85.1%), consistent with the known sex distribution of aortic disease, with a median age of 68 [64–73] years. The cohort was uniformly high-risk: 73.0% were American Society of Anesthesiologists (ASA) physical status III, with no ASA I or ASA V patients recorded. 643 procedures were elective and 50 were expedited (time-sensitive); no emergency cases were included. Variables differing significantly between patients with and without PPC included age ($P < 0.001$), COPD ($P < 0.001$), anesthesia type ($P = 0.001$), ASA physical status ($P = 0.031$), and predicted risks by Gupta RF ($P = 0.010$) and Gupta Pneumonia ($P < 0.001$); full distributions are presented in Table 1. Notably, the ARISCAT score as a continuous variable differed significantly between groups ($P = 0.018$), whereas its categorical risk strata (low, intermediate, high) did not ($P = 0.139$).

Outcomes

Overall outcome incidences are summarized in Table 2 and additional outcomes in Additional file 1: Table S2. The composite PPC endpoint occurred in 19.5% patients (n = 135); the Gupta RF endpoint in 8.7% (n = 60); and clinically diagnosed postoperative pneumonia in 9.1% (n = 63). Pneumonia (47% of PPC events) and acute respiratory failure (45%) together drove the majority of ARISCAT endpoint cases. Among the other PPC components, pleural effusion occurred in 6.3%, atelectasis in 3.2%, and bronchospasm in 3.0%. Unplanned reintubation, which constitutes a component of the Gupta RF endpoint, occurred in 7.1% of patients. Prolonged mechanical ventilation longer than 8 h occurred in 5.1%, and longer than 48 h in 1.7%.

Table 2. Incidence of primary and secondary postoperative outcomes in the study cohort (n = 693).

Discussion

Summary of findings

In a cohort of 693 patients undergoing OAAS, three widely used preoperative pulmonary risk scores — ARISCAT, Gupta RF, and Gupta Pneumonia — performed poorly for their intended primary endpoints. AUCs ranged from 0.563 to 0.604 for respective primary endpoints, with no significant difference between any score and its derivation outcome (all DeLong $P > 0.05$); the only cross-score finding was

that Gupta Pneumonia discriminated respiratory failure better than Gupta RF itself (P = 0.020). All three scores systematically underestimated risk: calibration-in-the-large was positive for each (+ 0.426, + 0.369, and + 0.837, respectively), and the calibration belt confirmed significant departure from the line of identity for all three scores (P ≤ 0.010).

Comparison with existing literature

The poor performance of ARISCAT in our cohort contrasts sharply with the c-statistics of approximately 0.90 in the development subsample and 0.88 in the internal validation subsample reported by Canet and colleagues [3], and around 0.80 in the multinational PERISCOPE external validation by Mazo and colleagues [6]. A central reason for this contrast is likely the case-mix of the present cohort. The derivation and earlier validation cohorts were broad non-cardiac surgical populations, in which a substantial proportion of patients had low baseline pulmonary risk; ARISCAT was designed to separate such low-risk patients from the smaller higher-risk subgroups. In OAAS, by contrast, baseline pulmonary risk is uniformly high — approximately 73% of our patients (503/693) fell into the intermediate ARISCAT category. The same compression was observed for Gupta RF, which produced only four unique predicted values across the entire cohort. When the predictor distribution is compressed in this way, even a well-derived model has little statistical room to discriminate; a phenomenon formally described by Vergouwe and Debray and colleagues [14, 15].

Performance in PERISCOPE differed markedly by region: discrimination was best in Western Europe (c-statistic 0.87, 95%CI 0.83–0.90) and lowest in the Eastern European (EE) subsample — which includes Poland, our study country — (c-statistic 0.76, 95%CI 0.72–0.80) with the highest PPC incidence (10.51%) and most pronounced calibration deterioration (slope 0.58) [6]. EE region remained an independent predictor of PPC even after ARISCAT adjustment, suggesting additional region-specific risk factors [6]. Our findings extend this pattern: even moderately good discrimination in EE general surgery was not preserved in the uniformly high-risk OAAS context.

The clinical utility data mirror this pattern. In the PERISCOPE EE subsample, the standard ARISCAT threshold of ≥ 26 points already yielded the poorest sensitivity of all subregions (54.3%) and a negative likelihood ratio of 0.56 [6]. In our OAAS cohort, the same threshold classified 82.5% of patients as at-risk, reducing specificity to 18.8% — effectively no better than the procedure type alone.

A similar pattern of case-mix compression leading to poor discrimination has been reported in other homogeneous high-risk populations. Kokotovic and colleagues externally validated ARISCAT in 585 patients undergoing major emergency abdominal surgery and observed systematic underestimation of PPC risk (predicted 24.9% versus observed 36.1%); despite this miscalibration, they reported good discrimination (AUC 0.83), indicating that underestimation alone does not preclude adequate score performance [7]. Siddiqui and colleagues, evaluating ARISCAT in cardiac surgery patients, demonstrated very high sensitivity (94.9%) but extremely poor specificity (4.65%), consistent with limited discriminatory utility in homogeneous high-risk populations [16]. The recent systematic review by Huang and colleagues concluded that no current model performs uniformly well across surgical specialties [17].

The Gupta Pneumonia score also underperformed substantially (AUC 0.604; O/E 2.17), despite a c-statistic of 0.86 in the original derivation cohort [5]. Two factors likely contribute: OAAS carries higher inherent pulmonary risk than the average National Surgical Quality Improvement Program (NSQIP) case, so predictions derived from a general surgical population will be systematically too low; and our pneumonia endpoint was not identical to the strict NSQIP definition, so some misclassification cannot be excluded. Notably, our observed incidences of respiratory failure (8.7%) and pneumonia (9.1%) are consistent with the 8%-12% (all pulmonary) and 5% (pneumonia) benchmark rates cited in Society for Vascular Surgery (SVS) guidelines for elective open AAA repair [8].

Clinical implications

From a clinical standpoint, our results have one principal practical implication. ARISCAT, the Gupta RF, and the Gupta Pneumonia scores, when applied at their originally proposed cut-offs, classify between three- and four-fifths of OAAS patients as high-risk (75–83%). At these thresholds, the scores essentially confirm what the clinician already knows from the procedure type itself, and offer little additional information for individualized decision-making. We would therefore caution against using these instruments to drive specific perioperative decisions based on the original cut-offs alone. For clinicians caring for this high-risk population, preoperative pulmonary risk assessment in OAAS should rely on multidisciplinary judgment informed by patient-specific factors — including functional reserve, pulmonary comorbidities, frailty [18], and anesthetic technique (general anesthesia alone was associated with higher PPC incidence; P = 0.001) — rather than on these generic scores.

Strengths

The principal strengths include the size and procedural specificity of the cohort: 693 patients undergoing the same high-risk procedure at a single tertiary center represent a large dedicated aortic validation dataset that enables direct head-to-head comparison of all three risk scores within a single, homogeneous surgical population. All variables required to compute the three scores were available with no missing data in the analytic cohort. The comprehensive calibration assessment goes beyond the discrimination-only approach common in earlier studies; original predicted probabilities were applied without refitting, giving a fair view of real-world score performance. Reporting followed the TRIPOD + AI 2024 statement.

Study limitations

Several limitations deserve emphasis. First, the study was retrospective and based on data from a single center, so its findings may not generalize to all OAAS practices, and especially not to centers with substantially different perioperative pathways or patient demographics, including sex composition, given the male predominance of our cohort (85%), which precluded formal fairness analyses by sex. Second, secondary analyses of non-pulmonary outcomes were exploratory and underpowered; findings are hypothesis-generating only. Finally, our pneumonia endpoint was the clinical diagnosis recorded in the medical record, which is not identical to the strict NSQIP definition used to derive the Gupta Pneumonia

score; some of the apparent miscalibration of that score may therefore reflect endpoint definition rather than model failure.

Future directions

Our findings support development of an OAAS-specific PPC prediction model. The present cohort (135 events, 693 patients) provides the basis for such work; procedure-specific variables — cross-clamp duration, clamp position, intraoperative blood loss, preoperative renal function — are plausible candidates not captured by existing scores. Multi-center external validation will be required before clinical implementation.

Conclusions

In a cohort of 693 patients undergoing OAAS, the ARISCAT, Gupta Respiratory Failure, and Gupta Pneumonia scores showed poor discrimination and significant miscalibration for their respective primary endpoints, with systematic underestimation of risk. None of these scores provides reliable individualized pulmonary risk stratification in this population. Aortic-specific prediction models are needed.

Abbreviations

AAA
abdominal aortic aneurysm
ARISCAT
Assess Respiratory Risk in Surgical Patients in Catalonia
ASA
American Society of Anesthesiologists
AUC
area under the receiver operating characteristic curve
CI
confidence interval
CITL
calibration-in-the-large
COPD
chronic obstructive pulmonary disease
DCA
decision curve analysis
EE
Eastern European
Gupta RF
Gupta Postoperative Respiratory Failure score

IQR
interquartile range
KDIGO
Kidney Disease:Improving Global Outcomes
NPV
negative predictive value
NSQIP
National Surgical Quality Improvement Program
O/E
observed-to-expected
OAAS
open abdominal aortic surgery
OSF
Open Science Framework
PAAD
peripheral arterial and aortic diseases
PPC
postoperative pulmonary complication
PPV
positive predictive value
ROC
receiver operating characteristic
RRT
renal replacement therapy
SVS
Society for Vascular Surgery
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence

Declarations

Ethics approval and consent to participate

Ethical approval for this study (No. OIL/KBL/13/2016) was provided by the Bioethics Committee of the District Medical Chamber in Krakow, Poland. The Committee waived the requirement for informed consent because of the retrospective design. The study was conducted in accordance with the latest version of the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to patient privacy regulations; anonymized data and the R analysis code are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any funding.

Authors' contributions

KWJ: conceptualization, formal analysis, writing – original draft, writing – review and editing. KP: conceptualization, supervision, writing – review and editing. MC: conceptualization, supervision, writing – review and editing. WS: conceptualization, supervision, writing – review and editing. KS: conceptualization, supervision, writing – review and editing. DS: conceptualization, data curation, project administration, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgements

During the preparation of this work the authors used generative AI tools - Claude Pro (Anthropic) and Gemini (Google, free version) - to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Table 2

Table 2. Incidence of primary and secondary postoperative outcomes in the study cohort (n = 693).

Outcome	n	%	Missing, n (%)
Primary outcomes			
PPC composite (ARISCAT endpoint)	135	19.5	0
Respiratory failure (Gupta RF endpoint)	60	8.7	0
Postoperative pneumonia (Gupta Pneumonia endpoint)	63	9.1	0
Components of PPC composite (ARISCAT)			
Acute respiratory failure	61	8.8	0
Postoperative pneumonia	63	9.1	0
Pleural effusion	44	6.3	0
Atelectasis	22	3.2	0
Bronchospasm	21	3.0	0
Aspiration pneumonitis	2	0.3	0
Pneumothorax	1	0.1	0
Mechanical ventilation and airway management			
Unplanned reintubation	49	7.1	1 (0.1)
Prolonged ventilation > 48 h	12	1.7	0
Prolonged ventilation > 8 h	35	5.1	0
Cardiac outcomes			
New-onset atrial fibrillation	27	3.9	0
Myocardial infarction	11	1.6	4 (0.6)
Pulmonary edema	13	1.9	0
Renal outcomes			
Acute kidney injury (KDIGO)	86	12.4	1 (0.1)
Renal replacement therapy	20	2.9	1 (0.1)
Mortality			
In-hospital mortality	49	7.1	0
30-day mortality	51	7.4	3 (0.4)
Systemic			

Outcome	n	%	Missing, n (%)
Primary outcomes			
Sepsis	24	3.5	0

References

1. Miskovic A, Lumb AB. Postoperative pulmonary complications. Br J Anaesth. 2017;118:317–34.
2. Fernandez-Bustamante A, Frendl G, Sprung J, Kor DJ, Subramaniam B, Martinez Ruiz R, et al. Postoperative pulmonary complications, early mortality, and hospital stay following noncardiothoracic surgery: a multicenter study by the Perioperative Research Network investigators. JAMA Surg. 2017;152:157–66.
3. Canet J, Gallart L, Gomar C, Paluzie G, Vallès J, Castillo J, et al; ARISCAT Group. Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010;113:1338–50.
4. Gupta H, Gupta PK, Fang X, Miller WJ, Cemaj S, Forse RA, et al. Development and validation of a risk calculator predicting postoperative respiratory failure. Chest. 2011;140:1207–15.
5. Gupta H, Gupta PK, Schuller D, Fang X, Miller WJ, Modrykamien A, et al. Development and validation of a risk calculator for predicting postoperative pneumonia. Mayo Clin Proc. 2013;88:1241–9.
6. Mazo V, Sabaté S, Canet J, Gallart L, Gama de Abreu M, Belda J, et al. Prospective external validation of a predictive score for postoperative pulmonary complications. Anesthesiology. 2014;121:219–31.
7. Kokotovic D, Degett TH, Ekeloef S, Burcharth J. The ARISCAT score is a promising model to predict postoperative pulmonary complications after major emergency abdominal surgery: an external validation in a Danish cohort. Eur J Trauma Emerg Surg. 2022;48:3863–7.
8. Chaikof EL, Dalman RL, Eskandari MK, Jackson BM, Lee WA, Mansour MA, et al. The Society for Vascular Surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. J Vasc Surg. 2018;67:2–77.e2.
9. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
11. Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. Stat Med. 2016;35:709–20.
12. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76.

13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
14. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68:279–89.
15. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971–980.
16. Siddiqui KM, Farooqui MH, Yousuf MS, Ali MA. ARISCAT and LAS VEGAS risk scores for predicting postoperative pulmonary complications after cardiac surgery: a cohort study. Ann Med Surg (Lond). 2024;86:3873–9.
17. Huang Z, Han Y, Zhuang H, Jiang J, Zhou C, Yu H. Prediction models for postoperative pulmonary complications: a systematic review and meta-analysis. Br J Anaesth. 2025;135:1415–27.
18. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ. 2005;173:489–95.

Figures

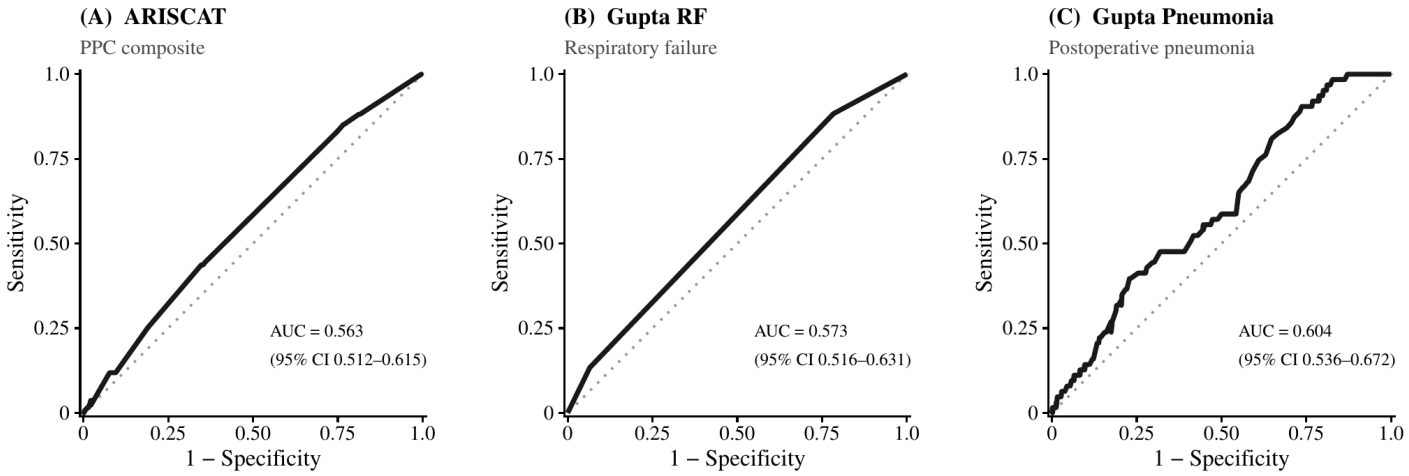


Figure 1

Receiver operating characteristic (ROC) curves of the three pulmonary risk scores. Each score is shown for its respective primary endpoint (n = 693). (A) ARISCAT for the postoperative pulmonary complication composite endpoint. (B) Gupta Respiratory Failure score for the respiratory failure endpoint. (C) Gupta Pneumonia score for the postoperative pneumonia endpoint. The diagonal dotted line represents no discrimination (AUC = 0.50). AUCs and 95% CIs were calculated by the DeLong method. AUC, area under the ROC curve; PPC, postoperative pulmonary complication. ARISCAT, Assess Respiratory Risk in Surgical Patients in Catalonia; Gupta RF, Gupta Postoperative Respiratory Failure score.

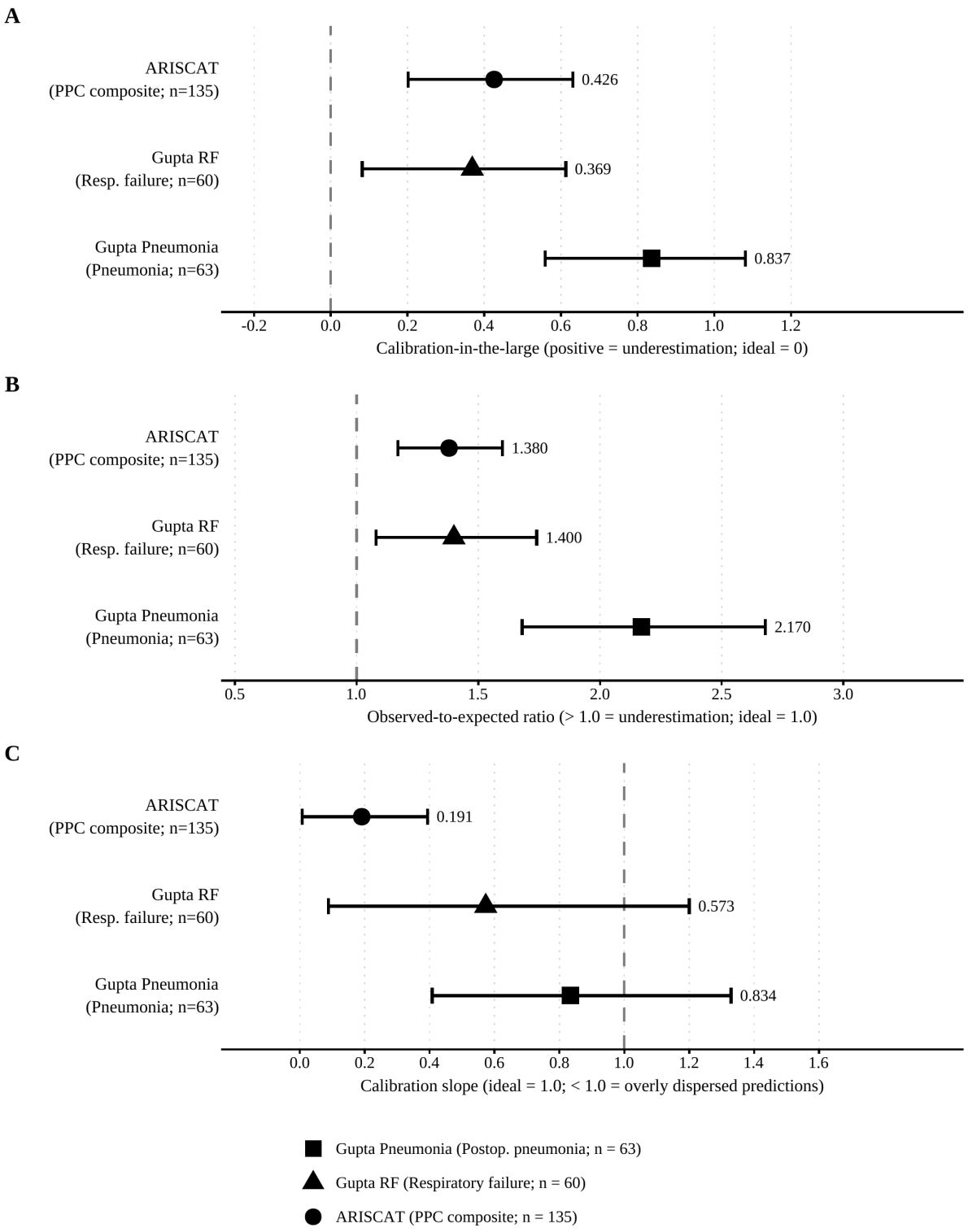
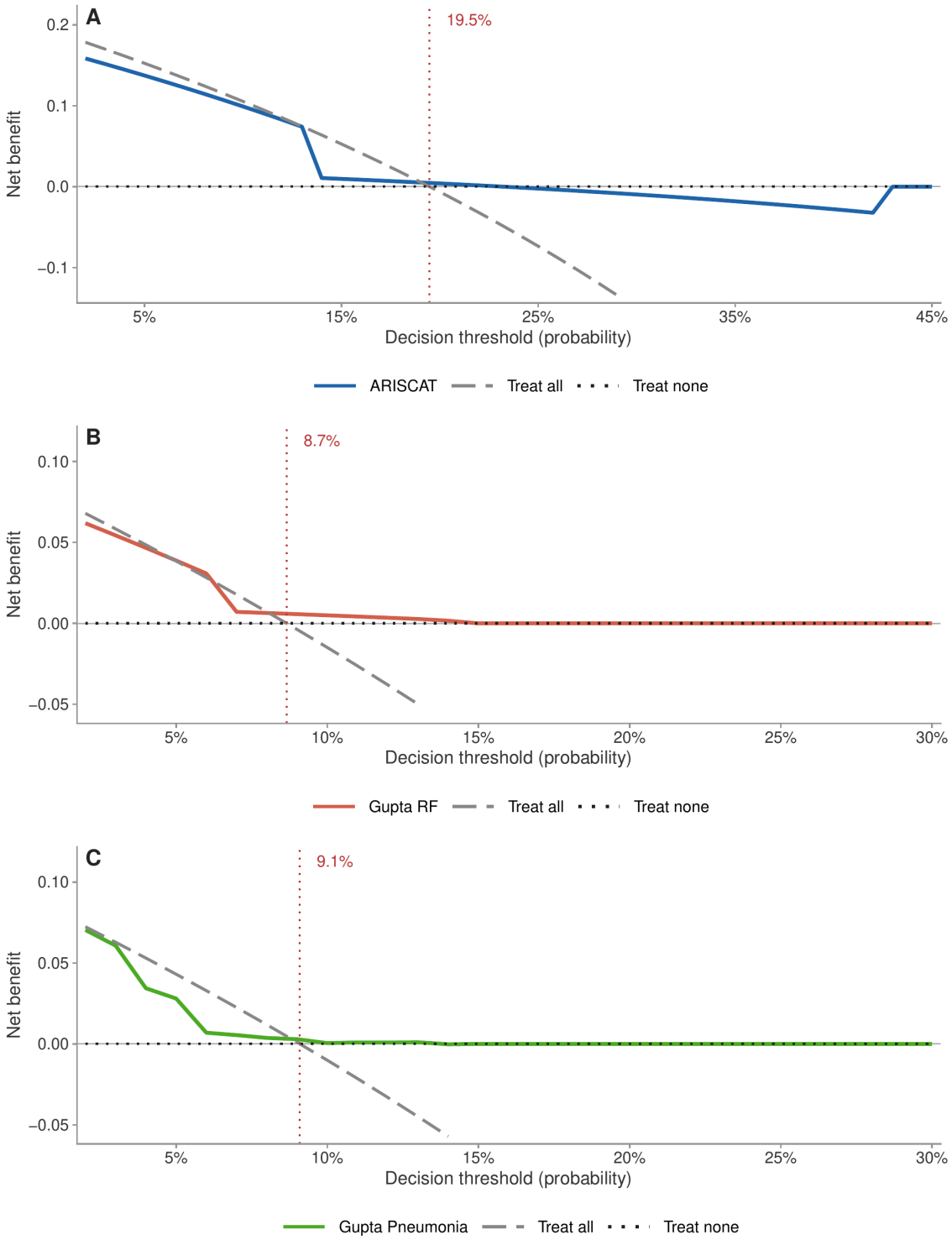


Figure 2

Calibration of three preoperative pulmonary risk scores applied to 693 patients undergoing open aortic surgery. Forest plots show point estimates (symbols) with bootstrap 95% confidence intervals (B = 2000) for calibration-in-the-large (panel A), observed-to-expected ratio (panel B), and calibration slope (panel C). The dashed vertical reference lines indicate ideal calibration (CITL = 0; O/E = 1.0; slope = 1.0). Calibration slope estimates for ARISCAT and Gupta RF are based on three and four unique predicted

probability values, respectively, and should be interpreted with caution. ARISCAT, Assess Respiratory Risk in Surgical Patients in Catalonia; CITL, calibration-in-the-large; O/E, observed-to-expected ratio; PPC, postoperative pulmonary complication. Gupta RF, Gupta Postoperative Respiratory Failure score.



Treat all: net benefit of classifying all patients as high−risk. Treat none: net benefit = 0 at all thresholds (reference). Red dotted vertical line: observed event

Figure 3

Decision curve analysis for the three pulmonary risk scores. (A) PPC composite endpoint, decision thresholds 5–50%. (B) Respiratory failure endpoint, decision thresholds 2–30%. (C) Postoperative pneumonia endpoint, decision thresholds 2–30%. Treat-all and treat-none reference strategies are shown for comparison. The red dotted vertical line in each panel indicates the observed event prevalence for the respective endpoint (A: 19.5%; B: 8.7%; C: 9.1%). DCA, decision curve analysis; PPC, postoperative pulmonary complication. ARISCAT, Assess Respiratory Risk in Surgical Patients in Catalonia; Gupta RF, Gupta Postoperative Respiratory Failure score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx